PROCEEDINGS, 51st Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 9-11, 2026
SGP-TR-230

Insights from Machine Learning Techniques for Heat Extraction Processes in Various Geothermal Resources

Edem Mensah and Mayank Tyagi

Louisiana State University, Baton Rouge, LA 70803

mtyagi@lsu.edu

Keywords: Geothermal Energy, Explainable Machine Learning, Non-negative Matrix Factorization, Self Organizing Map, Dimensionless Groups, Feature Extraction, Surrogate Model.

ABSTRACT

This study revisits the geothermal reservoir simulation datasets to provide insights through unsupervised as well as explainable machine learning (ML) techniques. Specifically, we have used Ansari and Hughes (2017) data that performed analysis of low-enthalpy hydrothermal reservoirs in the gulf coast regions of TX-LA border. Usefulness of self-organizing maps (SOM) and non-negative matrix factorization (NMF) in finding the important clusters among competing dimensionless groups through the operational life is demonstrated. Lastly, the results from the supervised ML methods – Random Forests (RF) and Artificial Neural Networks (ANNs) are compared and explained using SHAP features.

1. INTRODUCTION

Identifying promising reservoirs for geothermal energy production necessitates the use of predictive models. Various methods, such as inspectional analysis and statistical modeling, have been successfully employed to develop straightforward predictive models tailored to geothermal reservoir designs. In this study, we demonstrate the usefulness of unsupervised techniques such as self organizing maps and non-negative matrix factorization in identifying the important dimensionless groups governing the geothermal heat extraction at different times during the operational life. Further, the explainable machine learning such as SHAP features can delineate between supervised ML models such as Random Forests and ANNs.

2. PROBLEM DESCRIPTION AND DATASETS

Ansari and Hughes (2017) utilized inspectional analysis and statistical modeling to create simple predictive models for a regular line drive design. The regular line drive design (Figure 1) for geothermal reservoir management involves injecting cooler water at the up-dip side of the reservoir while extracting hot geofluid from the down-dip portion (Plaksina et al., 2011). This approach is preferred because it optimizes heat recovery by leveraging the natural thermal gradient of the reservoir. By injecting into the cooler sections and producing from the hotter regions, the design maximizes thermal efficiency and enhances overall energy extraction (Ansari et al., 2018).

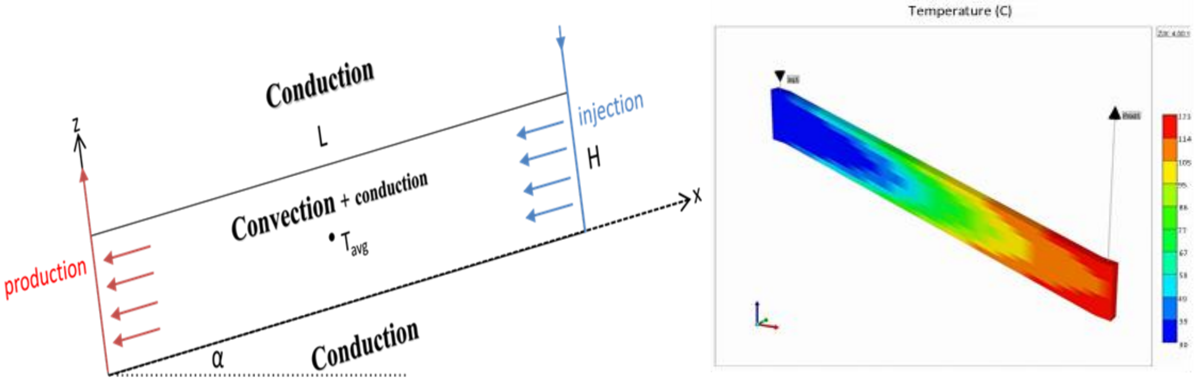


Figure 1: (Left) Description of Regular Line Drive System (Ansari 2016); (Right) Simulation Model (Ansari 2016)

Through inspectional analysis of the partial differential equations governing over 800 simulation models, fifteen dimensionless groups were identified (Ansari, 2016). These groups encapsulated the fundamental physics of the system, providing a simplified yet comprehensive representation of the reservoir dynamics (Table 1). The validity of these dimensionless groups was confirmed using models with similar groups but different dimensional parameters, ensuring their robustness and relevance in modeling two key responses: dimensionless production temperature and thermal recovery factor. By simulating different conditions and configurations, the models provided valuable insights into the behavior and efficiency of the geothermal systems, aiding in the identification of optimal reservoir candidates.

Table 1: Dimensionless Groups of Reservoir Input Parameters (Ansari, 2016)

Dimensionless Group	Feature
π_1	Total compressibility to fluid compressibility ratio
π_2	Total expansivity to fluid expansivity ratio
π_3	Fluid heat capacity ratio
π_4	Ratio of conductive to advective heat transfer (Fourier Number)
π_5	Fluid expansion due to average reservoir temperature
π_6	Effective aspect ratio
π_7	Dip angle group
π_8	geometric proportions of the geothermal reservoir
π_9	Buoyancy Number
π_{10}	Thermal Peclet number
π_{11}	Thermal diffusivity ratio
π_{12}	Thermal diffusivity ratio (secondary)
π_{13}	Fluid compression as a result of reservoir pressure
π_{14}	Temperature ratio
π_{15}	Temperature distribution

Leveraging ML methods, we propose to revisit the data from Ansari & Hughes (2017), to reveal significant relationships between production temperature and geophysical input and operational parameters during the operational life of geothermal reservoir. Following research tasks were undertaken: i) Non-negative Matrix Factorization (NMF) and Self Organizing Maps (SOM) to identify clusters/patterns among dimensionless temperature, and across groups, and ii) Explainable supervised machine learning models using SHAP values to identify features importance and forecast productivity metrics of geothermal reservoir.

3. OVERVIEW OF MACHINE LEARNING ALGORITHMS

A brief overview of the select ML algorithms is provided in the following subsections and Appendix for the sake of completeness.

3.1 Non-negative Matrix Factorization (NMF)

This algorithm combines a simple matrix factorization with a tailored version of k-means clustering (Ahmed et al. 2021). Non-Negative Matrix Factorization (NMF) belongs to a group of linear decomposition methods that includes Principal Component Analysis (PCA) and Independent Component Analysis (ICA). While all these techniques reduce data dimensionality by learning a linear representation, NMF stands out due to its unique characteristics. The primary advantage of NMF is its interpretability. Unlike PCA and ICA, which may obscure the meaning of their components, NMF's additive and parts-based nature ensures that its components are meaningful and interpretable (Ahmed et,.al 2020). Additionally, NMF excels at uncovering latent features, enabling the discovery of dominant signals and attributes in complex datasets. By integrating k-means clustering with NMF, the algorithm further enhances its capability to group data effectively based on the latent features revealed through the factorization process.

In this study, NMF is employed to decompose a non-negative data matrix into two smaller matrices, facilitating the discovery of hidden patterns or latent structures within the data. Meanwhile, the customized k-means clustering groups data points based on their similarity, enhancing the interpretability of the decomposed components. The K parameter, the number of distinct k means cluster or signatures can range from 2 to the maximum number of locations or attributes. For each value of K, NMF is executed 1,000 times, and the solution with the smallest reconstruction error is selected as the optimal result (Vesselinov et al. 2020). Subsequently, the 1,000 matrices are clustered into groups within a dimensional space. The quality of clustering is evaluated using silhouette widths, computed based on a cosine similarity norm. Optimal signals are characterized by a combination of minimal reconstruction error and high silhouette width, reflecting their precision and coherence in representing latent data structures. For more details see Appendix.

3.2 Self Organizing Maps (SOM)

SOM, developed by Kohonen, are unsupervised machine learning algorithms that project high-dimensional data onto a lower-dimensional grid while maintaining topological relationships (Yin, 2008). The standard SOM architecture utilizes a two (2) layered network, an input layer and an output layer arranged into a two-dimensional grid (see Appendix). SOMs are particularly effective in clustering, visualizing, and analyzing complex datasets, making them suitable for hydrological modeling, soil characterization, and other environmental studies (Lee & Kim, 2021). Two primary metrics are used to evaluate the quality of a trained SOM: quantization error (QE) and topographic error (TE) extraction (Bacao et al., 2005). QE measures the average distance between each data vector and its corresponding BMU, serving as an indicator of how well the SOM represents the data distribution. A lower QE suggests a better fit between the data vectors and the neuron vectors. TE, on the other hand, evaluates the topology preservation of the SOM by measuring the average distance between the BMU and the second-best matching neuron for each data vector. A lower TE indicates that the SOM effectively replicates the topology of the data in its original space. Algorithms such as k-means or DBSCAN can refine clustering by analyzing neuron weight vectors. Techniques like label propagation and threshold-based methods further enhance clustering by leveraging SOM's topology (Pham, et al., 2017). Key considerations include optimizing SOM parameters and selecting the appropriate number of clusters. Applications span various domains, including geophysical analysis, customer segmentation, image classification, and healthcare, showcasing SOM clustering as an effective method for uncovering complex data patterns. For more details see Appendix.

3.3 SHapley Additive exPlanations (SHAP)

SHAP analysis is applied to each supervised machine learning model to enhance explainability. SHAP values provide a quantitative measure of the contribution of each dimensionless reservoir feature to the model's predictions, allowing for a deeper understanding of feature importance and their respective impacts on temperature predictions (Lundberg and Lee, 2017). By analyzing SHAP distributions, key insights such as Feature Contribution, Positive and Negative Effects, and Cluster-Specific Variability, emerge. SHAP analysis enhances model transparency, supporting better decision-making in geothermal and petroleum applications while improving confidence in machine learning-driven predictions. For more details see Appendix.

3.4 Random Forests (RF)

RF is a highly effective machine learning algorithm that uses an ensemble of decision trees to improve predictive accuracy and reduce generalization errors. By combining the outputs of multiple decision trees, Random Forest achieves a single, reliable prediction, leveraging the "wisdom of crowds" principle. This approach makes it well-suited for handling both regression and classification tasks, offering robust and interpretable results. Compared to other ensemble methods like XGBoost, Random Forest constructs decision trees independently and averages their predictions, ensuring simplicity and resilience in varied applications. Key parameters such as pressure, depth, and thermal conductivity could be ranked based on their influence on temperature fluctuations and well behavior. This feature ranking enables targeted decision-making and facilitates operational improvements by focusing on the most impactful variables. Additionally, Random Forest is highly effective at detecting anomalies in well temperature profiles.

3.4 Artificial Neural Networks (ANN)

ANNs are a category of machine learning models inspired by the structure and functioning of the human brain. These models are capable of learning from data and applying that knowledge to make predictions or decisions. ANNs have found extensive applications across various domains, including oil and gas exploration and development. Key parameters that play a critical role in the design, performance, and generalization of a Deep Neural Network, are: Number of Layers, Number of Units per Layer, Dropout Rate, Learning Rate, Batch Size, Loss, and Activation functions. Careful tuning and selection of these hyperparameters are essential for building robust and high-performing ANNs tailored for wide range of engineering applications. For more details see Appendix.

4. RESULTS AND DISCUSSIONS

4.1 SOM

After training, the weights of the SOM neurons are extracted and flattened. These weights serve as input to the K-Means clustering algorithm, which groups neurons into distinct clusters (Table 2). An iterative process was used to identify the best number of clusters, based on silhouette scores. In this study, four 4 distinct clusters were identified. As shown below, the central red region (Cluster 3: Late Regime) stands out as a dominant cluster, reflecting a significant grouping of data points with similar characteristics mapped to this area. Surrounding this central cluster are additional distinct clusters, including the green region (Cluster 1: Late Intermediate Regime) and the purple regions (Cluster 2: Early Intermediate Regime). These regions highlight areas where data points exhibit varying properties. Additionally, the blue region (Cluster 4: Early Regime) represents another distinct and unique grouping within the SOM grid. The large green cluster (Cluster 1) occupies a substantial portion of the grid, indicating that a significant part of the dataset shares common features. Conversely, the smaller purple clusters (Cluster 2) might represent specialized subsets of the data, possibly corresponding to unique patterns or outliers. The boundaries separating these clusters represent regions where the SOM has effectively captured transitions in the feature space. These boundaries, consistent with high-distance areas in the U-Matrix, emphasize the distinct separations between groups within the dataset.

Overall, the cluster distribution across the SOM grid reflects a non-uniform structure, comprising several well-defined groups. The compact configuration of the red cluster (Cluster 3) suggests a dense collection of similar data points, while the more dispersed arrangement of other clusters highlights variability and diversity within the dataset. In Cluster 1 (Late Intermediate Regime) and Cluster 2 (Early Intermediate Regime), the ratio of injection temperature to the average reservoir temperature (π14) shows strongest positive correlations with dimensionless temperature values, suggesting that the injection temperature of fluid (water) has a high influence on the production temperature in this regime. Effective aspect ratio (π6) and reservoir geometry (reservoir length/ reservoir height) (π8) also exhibit moderate positive correlations, but the values are less pronounced compared to π14. It can also be observed that π5, the fluid expansion due to average reservoir temperature and π10 the peclet number dominates with strong negative correlations. Distinctly different patterns emerge with stronger correlations for reservoir dip angle (π7) and the temperature gradient (i.e. temperature distribution, π15) in the later operational period of the reservoir (I.e. the late regime). However, that π5, the fluid expansion due to average reservoir temperature, π9, the buoyancy number and π10 the peclet number, indicate strong negative correlation, implying an inverse relationship. Finally, Cluster 4 (early regime) sees π14 continuing to dominate with strong positive correlations while π6 and π8 maintain moderate positive trends, as also observed in the intermediate regimes.

Cluster	Dimensionless Temperature Ranges
Cluster 1 (Late Intermediate Regime)	$TD_{19} – TD_{27}$
Cluster 2 (Early Intermediate Regime)	$TD_{11} – TD_{18}$
Cluster 3 (Late Regime)	$TD_{28} – TD_{40}$
Cluster 4 (Early Regime)	$TD_{1} – TD_{10}$

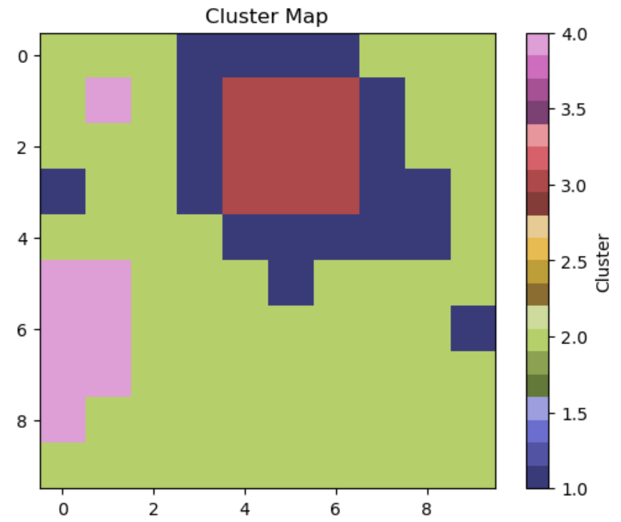


Table 2: SOM Clusters Ranges and SOM Cluster Map.

4.2 NMF

From figure 2, Signatures 2 and 4 (dominate in the early regime, reflecting temperature dynamics that are most pronounced in the initial stages of the system. These signatures highlight the influence of conditions prevalent in the early stages, such as rapid thermal changes and localized effects, which are often critical for understanding early-stage processes. The intermediate regime is characterized by Signatures 3 and 5, which capture transitional dynamics that occur as the system progresses from early-stage influences toward a more stable phase. These signatures underline the gradual adjustments in temperature and energy transfer, reflecting more distributed and evolving patterns in this phase. Signature 1 is associated with the late regime, demonstrating consistent contributions across the latter part of the system's progression. This signature emphasizes stability and sustained behavior, often indicative of equilibrium or long-term patterns in the system's thermal characteristics.

After identifying the temperature signatures, there is a need to estimate the correlation between these signatures and the measured and weighted reservoir parameters and features as shown in SOM analysis. Replicating the same procedure, Pearson's correlation computed. The heatmap below figure 2 presents the correlation between various measured and weighted reservoir parameters and features and the five hidden temperature signatures identified using NMF. The color intensity reflects the strength and direction of the correlation, ranging from strongly positive (red) to strongly negative (blue).

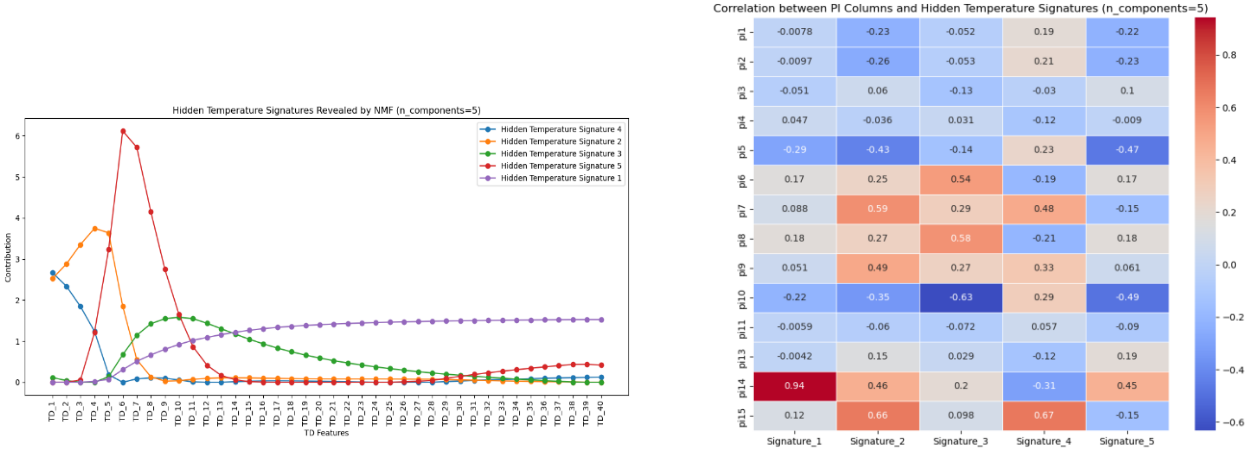


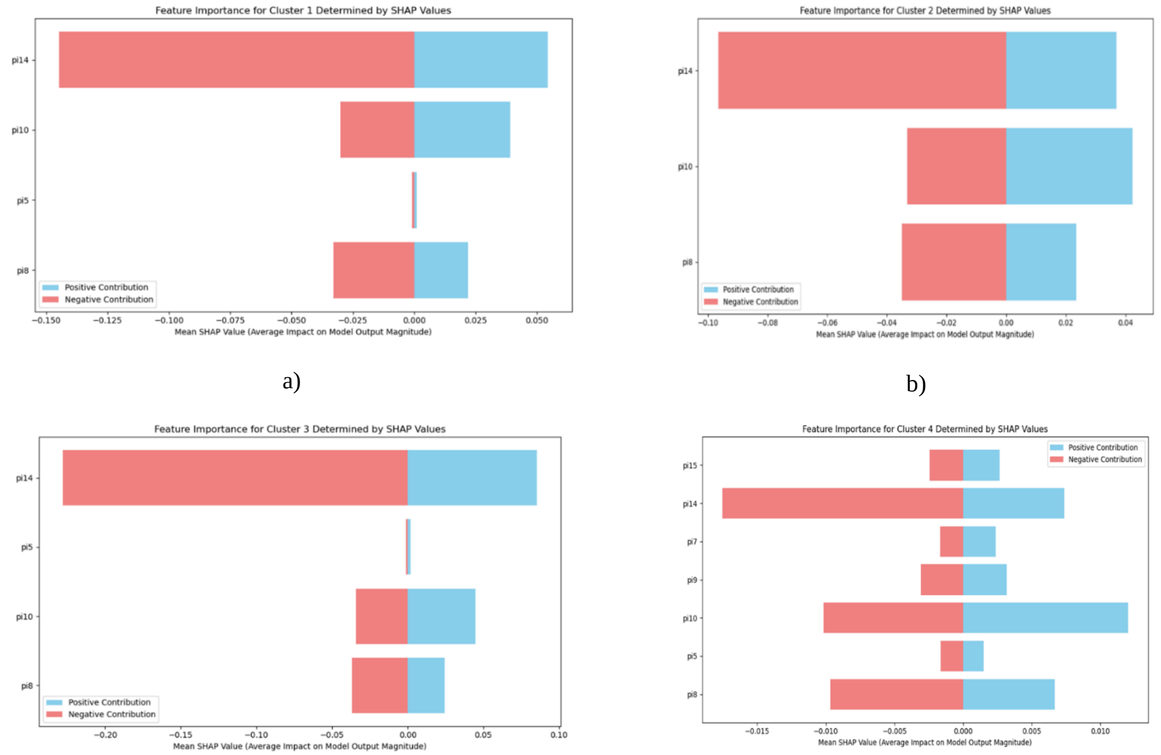
Figure 2 TD Feature Contribution, Dimensionless Feature vs Hidden Temperature Signatures as Correlation Heatmap

Signature 1, associated with the late regime, is highly correlated with π_{14} (Temperature ratio; 0.94), indicating a strong association with late-stage performance indicators. Signature 2, representing the early regime, shows strong positive correlations with π_{15} (Temperature distribution; 0.66) and moderate correlations with π_{7} (Dip angle group; 0.59) and π_{9} (Geometric proportions of the geothermal reservoir; 0.49). In contrast, it has negative correlations with π_{5} (Fluid expansion due to average reservoir temperature; -0.43) and π10 (Thermal Peclet number; -0.35), highlighting potential trade-offs in early-stage performance dynamics. Signature 3, corresponding to the early intermediate regime, has moderate positive correlations with π_6 (Effective aspect ratio; 0.54) and π_8 (Geometric proportions of the geothermal reservoir; 0.58), indicating their importance in intermediate-stage dynamics. Signature 4, also tied to the early regime, correlates positively with π_7 (Dip angle group; 0.48) and π_9 (Geometric proportions of the geothermal reservoir; 0.33). However, it has a strong negative correlation with $\pi 10$ (Thermal Peclet number; -0.63), suggesting contrasting influences in this regime. Finally, Signature 5, linked to the late intermediate regime, exhibits positive correlations with π_{15} (Temperature distribution; 0.67) and π_{14} (Temperature ratio; 0.45), underlining their significance in this phase. It has negative correlations with π_5 (Fluid expansion due to average reservoir temperature; -0.47) and $\pi 10$ (Thermal Peclet number; -0.49), indicating reduced relevance in this regime.

In summary, π_{14} (Temperature ratio) and π_{15} (Temperature distribution) emerge as critical indicators for late and late-intermediate regimes, with strong positive correlations in Signatures 1 and 5. Early regime dynamics, reflected in Signatures 2 and 4, are influenced by π_7 (Dip angle group) and π_9 (Geometric proportions of the geothermal reservoir), while π_{10} (Thermal Peclet number) has a contrasting negative association. Intermediate regime signatures, represented by Signatures 3 and 5, highlight the relevance of π_6 (Effective aspect ratio) and π_8 (Geometric proportions of the geothermal reservoir). This correlation analysis provides valuable insights into the key performance indicators that contribute to specific temperature regimes, offering a framework for optimizing system behavior and understanding regime-specific dynamics.

4.3 RF

From figure 3, the SHAP value bar plots illustrate the importance of key Profiling Indicators (PIs) in the Random Forest model predictions for each cluster. In Cluster 1, the most impactful PIs were π_{14} (Temperature ratio) and π_{10} (Thermal Peclet number), with substantial contributions from π_5 (Fluid expansion due to average reservoir temperature) and π_8 (Geometric proportions of the geothermal reservoir). These PIs collectively influenced the model's ability to predict temperature behaviors accurately, with positive and negative contributions well-balanced. For Cluster 2, a smaller set of PIs, including π_{14} (Temperature ratio), π_{10} (Thermal Peclet number), and π_8 (Geometric proportions of the geothermal reservoir), demonstrated significant contributions. These features captured the essential relationships required for modeling temperature variations effectively in this cluster. The SHAP values for these PIs were consistent, highlighting their reliability as predictors. In Cluster 3, a broader set of PIs, led by π_{14} (Temperature ratio) and π_5 (Fluid expansion due to average reservoir temperature), provided the model with a comprehensive understanding of temperature dynamics. π_{10} (Thermal Peclet number) and π_8 (Geometric proportions of the geothermal reservoir) also had notable impacts, while contributions from π_6 (Effective aspect ratio) and π_9 (Buoyancy Number) enriched the model's predictive capability. This combination of PIs enabled the model to capture intricate temperature-related patterns specific to Cluster 3. Finally, Cluster 4 revealed the importance of π_{15} (Temperature distribution), π_{14} (Temperature ratio), and π_7 (Dip angle group), with additional contributions from π_9 (Buoyancy Number) and π_{10} (Thermal Peclet number). π_5 (Fluid expansion due to average reservoir temperature) and π_8 (Geometric proportions of the geothermal reservoir) also played critical roles in refining the model's accuracy. The SHAP analysis for this cluster highlighted a well-distributed impact of positive and negative contributions among the selected PIs.



a)

b)

c)

d)

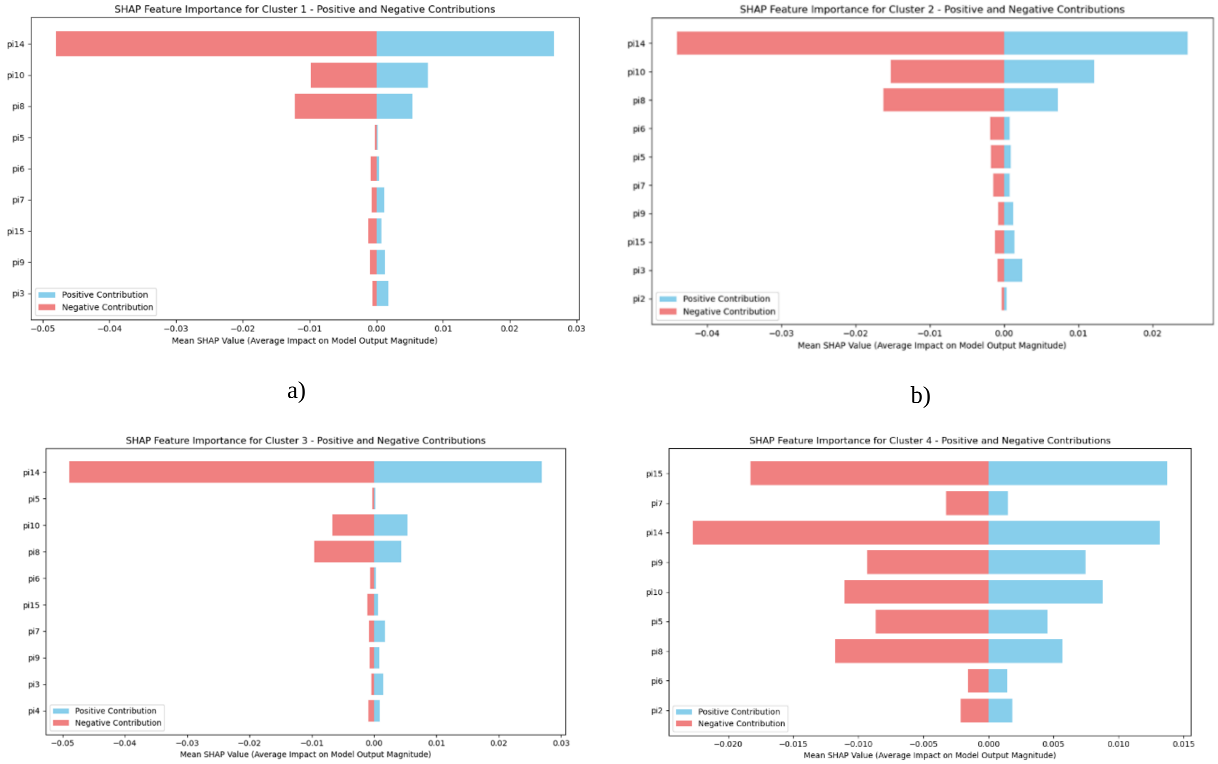
Figure 3: Explainable SHAP Analysis for Random Forest Models.

Across all clusters, π_{14} (Temperature ratio) consistently stood out as a critical profiling indicator (PI), demonstrating its universal importance in temperature-related predictions. While the number and type of impactful PIs varied among clusters, the contributions of each feature aligned with the unique characteristics of their respective regimes. These SHAP value plots reinforce the effectiveness of the selected PIs and the robustness of the Random Forest models in capturing complex relationships within clustered datasets.

The performance (Appendix Table A.1) of the Random Forest models across all clusters was evaluated using key metrics, including Root Mean Squared Error (RMSE) and R^2 scores for both training and testing datasets. These metrics highlight the models' accuracy and ability to generalize well to unseen data. While the training RMSE was the lowest among all clusters at 0.0040, the testing RMSE was 0.0050. However, the R^2 scores were relatively lower, at 0.8570 for training and 0.8166 for testing, indicating that while the model performed well during training, its generalization to unseen data was comparatively less robust.

4.4 ANN

Figure 4 shows the results of SHAP analysis employed to interpret the contributions of key dimensionless groups (PIs) in the ANN models. In Cluster 1, π_{14} (Temperature ratio) was the most impactful variable. This highlights the dominant influence of thermal contrasts in the reservoir. The next most influential features were π_{10} (Thermal Peclet number), associated with the balance of advective and conductive heat transport, and π_8 (geometric proportions of the geothermal reservoir), emphasizing structural influences on thermal flow behavior.



a)

b)

c)

d)

Figure 4: Explainable SHAP Analysis for ANN Models.

Similarly, in Cluster 2, π_{14} (Temperature ratio) remained the top contributing feature, reaffirming the importance of temperature gradients. π_{10} (Thermal Peclet number), and π_8 (geometric proportions of the geothermal reservoir) also played prominent roles, further underscoring the significance of heat transport dynamics and reservoir geometry in shaping thermal responses. In Cluster 3, π_{14} (Temperature ratio) continued to dominate, but the impact of π_5 (Fluid expansion due to average reservoir temperature) was also notable. This suggests an increased relevance of thermal fluid dynamics and fluid-rock interaction mechanisms. π_{10} (Thermal Peclet number) again emerged as a supporting contributor, reinforcing the persistent influence of convective transport. For Cluster 4, a more distributed feature importance profile emerged. While π_{15} (Temperature distribution), π_{14} (Temperature ratio) were still significant, other features like π_7 (Dip angle group), π_9 (Buoyancy Number), and π_{10} (Thermal Peclet number) indicates a more complex interplay of geometry, stratigraphy, and buoyancy-driven effects (Figure 4). This broader influence spectrum points to a more heterogeneous reservoir regime.

5. CONCLUSIONS

The methodological workflow established in this study spanning unsupervised clustering, latent feature recognition, and supervised prediction with explainability—offers a transferable framework for advanced geothermal reservoir modeling. This integrated approach not only enhances predictive accuracy but also deepens the interpretability of results, offering a powerful decision-making tool for geothermal exploration and production optimization. Following conclusions are made:

i. Temperature Ratio (π_{14}) emerged as the most influential variable across all clusters, reaffirming its importance as a key descriptor of thermal performance in geothermal systems.

ii. The combination of SOM and NMF proved highly effective in segmenting the reservoir behavior and extracting physically meaningful latent features.

iii. Supervised learning models, particularly ANNs, demonstrated robust predictive capabilities, with SHAP analysis offering a transparent explanation of model behavior.

iv. Each thermal regime was influenced by different subsets of dimensionless groups, with early regimes being more sensitive to structural and geometric properties, while late regimes were governed by thermal dynamics.

ACKNOWLEDGMENT

Author (EM) would like to thank the Craft & Hawkins Department of Petroleum Engineering for the financial support as well as the Ghana Education Trust Fund (GETFund) for their generous funding during the early years of his graduate studies. Authors would like to thank Dr. Esmail Ansari for the dataset and Prof. R.G. Hughes (LSU) for discussions on explainable ML results. Authors acknowledge the support from the LSU Center for Computation and Technology (CCT) and LSU High-Performance Computing (HPC) resources during this research.

APPENDIX – OVERVIEW OF SELECTED MACHINE LEARNING TECHNIQUES

Self Organizing Maps: Prior to training a SOM, the dataset must be normalized to ensure all variables lie within a similar range. SOMs treat each data point as an n-dimensional vector, with each component of the vector representing a unique data type. The SOM neural network consists of computational units, called neurons, typically arranged in a two-dimensional grid. Each neuron is represented as a vector, with components corresponding to the same features as the data vectors (Yin, 2008). These neuron vectors, often referred to as weights or reference vectors, are initialized and adjusted during the training process.

Initially, the components of the neuron vectors are set randomly to avoid introducing bias into the training process. In some SOM implementations, neuron values may be initialized using prior information, but this approach is not always employed to ensure unbiased learning (Kohonen, 1982).

The training process involves iterative modifications to the neuron vectors. Each data vector is sequentially introduced to the SOM. The neuron vector most similar to the data vector is identified and designated as the Best Matching Unit (BMU). This similarity is typically measured using a distance metric, such as Euclidean distance. Graphical representations of the trained SOM include component plots and the unified distance matrix (U-matrix). Component plots visualize the distribution of a single feature by coloring the neurons based on the value of one component of their vectors. These plots provide insights into the normalized distribution of features in the data space. The U-matrix visualizes the organization of the SOM by coloring neurons according to their similarity to immediate neighbors. Regions of high similarity indicate well-defined clusters, while areas of low similarity suggest boundaries between clusters. Once the SOM is trained, each data vector is mapped to its BMU, effectively clustering the data. Neurons can be interpreted as representatives of small groups of data points. According to Yin, 2008, the iterative training process can be mathematically expressed as:

$$W_{i,j}(t+1)=W_{,i,j}(t)+\alpha(t)(x_i(t)-W_{i,j}(t))$$

where:

$W_{i,j}(t))$ represents the weight of neuron j at time step t.

$\alpha(t)$ is the learning rate at time t.

$x_i(t)$ is the input vector at time t.

Over time, the SOM organizes itself such that similar data points map to neighboring neurons in the grid. This property makes SOMs particularly effective for visualizing high-dimensional data in two-dimensional spaces and identifying underlying patterns or clusters in the dataset (Yin, 2008).

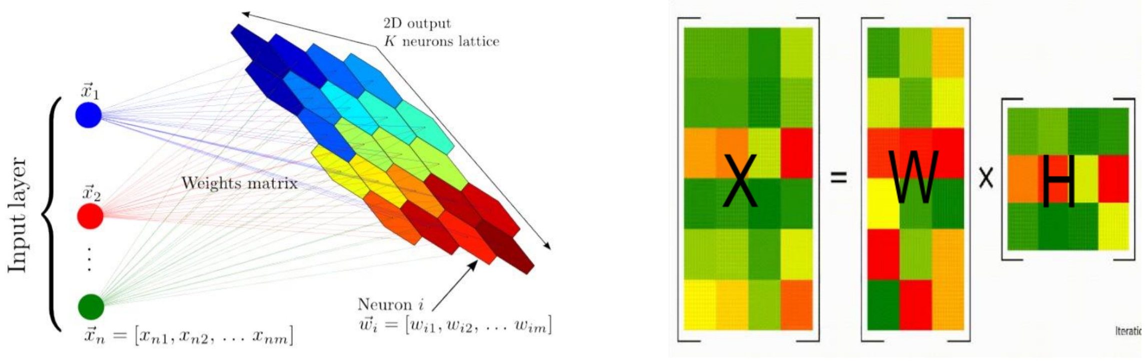


Figure A.1: Self-Organizing Map Architecture (https://www.latentview.com/blog/self-organizing-maps/), NMF(K) Schematic (Ahmed et,.al 2020)

Non-Negative Matrix Factorization: Given a non-negative data matrix, each column represents a variable or sample vector. The matrix has rows corresponding to locations and columns corresponding to attributes. Non-Negative Matrix Factorization (NMF) decomposes into two smaller non-negative matrices and where is a user-defined parameter representing the number of dimensions or latent features. According to Ahmed et,.al (2020), this decomposition is achieved by minimizing the following loss function:

$$\boldsymbol{L = \| X - WH^T \|_F^2}$$

Where $\mathcal{L}$ denotes the Frobenius norm. The matrix serves as a basis matrix, optimized to provide a linear approximation. The matrix serves as a basis matrix, optimized to provide a linear approximation. Since only a limited number of basis vectors are used to approximate the entire dataset, these vectors effectively capture the latent structures within.

Shapley Value: The Shapley value is a concept in game theory used to determine the contribution of each player in a coalition or a cooperative game (Lundberg and Lee, 2017). Assume teamwork is needed to finish a project. The team, T, has members. Knowing that the contribution of team members during the work was not the same, how can we distribute the total value achieved through this teamwork, v = v(T), among team members? Shapley value, $\phi_m(\mathrm{v})$, is the fair share or payout to be given to each team member. The $\phi_m(\mathrm{v})$ is defined as:

$$\boldsymbol{\phi_m(v) = \frac{1}{p}\sum \frac{[v(S \cup \{m\}) - v(S)]}{(p-1,\ k(s))}}$$

For a given member, $T = \{1,2,3,\ldots,p\}$, the summation is over all the subsets S, of the team, $T = \{1,2,3,\ldots,p\}$, that one can construct after excluding m. In the $k(S)$ above formula, is the size of $S, v(S)$, is the value achieved by subteam S, and $v(S \cup \{m\})$ is the realized value after m joins S.

Random Forests: Random Forest builds on the foundation of Classification and Regression Trees (CART), employing regression trees for prediction tasks. The algorithm recursively splits internal nodes to identify optimal features and splitting points, continuing until the termination condition. The algorithm is based on the following hyperparameters such as: n_estimators: The number of trees in the forest, max_depth: The maximum depth of each tree, balancing complexity and overfitting, min_samples_split: The minimum number of samples required to split an internal node, min_samples_leaf: The minimum number of samples in a leaf node, max_features: The number of features considered at each split.

Table A.1: Random Forest Training and Test Model Performance in this study

CLUSTER	Training RMSE	Testing RMSE	Training R²	Testing R²
Cluster 1 (Late Intermediate Regime)	0.0086	0.0111	0.9728	0.9580
Cluster 2 (Early Intermediate Regime)	0.0093	0.0128	0.9574	0.9236
Cluster 3 (Late Regime)	0.0074	0.0094	0.9815	0.9712
Cluster 4 (Early Regime)	0.0040	0.0050	0.8570	0.8166

Artificial Neural Networks: ANNs are composed of three fundamental layers: the input layer, one or more hidden layers, and the output layer. The input layer accepts data, typically formatted as a numerical vector (Benti, et al 2023). The hidden layers process the data through interconnected neurons that perform nonlinear transformations. Finally, the output layer produces the network's prediction or classification. Each neuron in the network receives input either directly from the input layer or from neurons in previous layers. Every connection between neurons is assigned a weight, representing the strength of influence between them. The neuron calculates a weighted sum of its inputs and applies an activation function σ_j to introduce nonlinearity (Wang, et al 2022). This output is then passed on to subsequent neurons, continuing through the network until the final output is generated in the output layer.

$$\mathrm{z_i = \sum_{j=1}^{n} \sigma_j\left(w_{ij}x_i + b_j\right)}$$

Table A.2: ANN Training and Test Model Performance in this study

CLUSTER	Training RMSE	Testing RMSE	Training R²	Testing R²
Cluster 1 (Late Intermediate Regime)	0.01995	0.02238	0.97263	0.97378
Cluster 2 (Early Intermediate Regime)	0.02812	0.02914	0.95474	0.96024
Cluster 3 (Late Regime)	0.01519	0.01644	0.98288	0.98501
Cluster 4 (Early Regime)	0.03486	0.03872	0.94168	0.94013

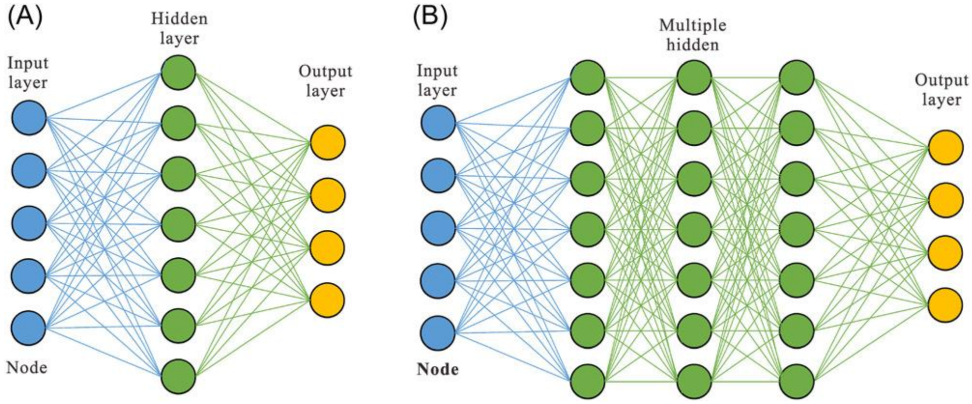


Figure A.2: Shallow neural network diagram and (B) deep neural network diagram (Wang, et al 2022).

REFERENCES

Verma, A., and Pruess, K.: Enhancement of Steam Phase Relative Permeability Due to Phase Transformation Effects in Porous Media, Proceedings, 11th Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, CA (1986). <Reference Style>

Ahmmed, B., Karra, S., Vesselinov, V. V., and Mudunuru, M. K. (2021). Machine learning to discover mineral trapping signatures due to CO2 injection. International Journal of Greenhouse Gas Control, 109, 103382.

Ahmmed, B., Vesselinov, V., and MK, M. (2020). Non-negative matrix factorization to discover dominant attributes in Utah FORGE Data. Geothermal Resources Council, Reno, NV.

Ahmmed, B., Vesselinov, V. V., and Mudunuru, M. K. (2021). Non-negative matrix factorization to discover dominant attributes in Utah FORGE Data (No. LA-UR-20-26171). Los Alamos National Laboratory (LANL), Los Alamos, NM (United States).

Alqahtani, A., Alsubai, S., Sha, M., and Dutta, A. K. (2025). Revolutionizing ALS Assessment: XGBoost Classification with Progressive Entropy Weighted-based Focal Loss on Gene Sequences. Journal of Disability Research, 4(1), 20240119.

Ansari, E., Hughes, R., and White, C. D. (2018). Modeling a new design for extracting energy from geopressured geothermal reservoirs. Geothermics, 71, 339-356.

Ansari, E., Hughes, R., and White, C. D. (2017). Statistical modeling of geopressured geothermal reservoirs. Computers and Geosciences, 101, 36-50.

Ansari, E., and Hughes, R. (2016). Response surface method for assessing energy production from geopressured geothermal reservoirs. Geothermal Energy, 4, 15.

Ansari, E. (2016). Mathematical scaling and statistical modeling of geopressured geothermal reservoirs [Doctoral dissertation, Louisiana State University and Agricultural and Mechanical College]. LSU Digital Commons. https://digitalcommons.lsu.edu/gradschool_dissertations/1234

Bação, F., Lobo, V., and Painho, M. (2005). Self-organizing maps as substitutes for k-means clustering. In International Conference on Computational Science (pp. 476-483). Berlin, Heidelberg: Springer Berlin Heidelberg.

Benti, N. E., Chaka, M. D., and Semie, A. G. (2023). Forecasting renewable energy generation with machine learning and deep learning: Current advances and future prospects. Sustainability, 15(9), 7087.

Brown, S., Rodi, W. L., Gu, C., Fehler, M., Faulds, J., Smith, C. M., and Treitel, S. (2022). Bayesian Neural Networks for Geothermal Resource Assessment: Prediction with Uncertainty. arXiv preprint arXiv:2209.15543.

Chen, T., and Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Dalal, S., Lilhore, U. K., Faujdar, N., Simaiya, S., Agrawal, A., Rani, U., and Mohan, A. (2024). Enhancing thyroid disease prediction with improved XGBoost model and bias management techniques. Multimedia Tools and Applications, 1-32.

Ertekin, T., and Sun, Q. (2019). Artificial intelligence applications in reservoir engineering: a status check. Energies, 12(15), 2897.

Han, J. X., Xue, L., Wei, Y. S., Qi, Y. D., Wang, J. L., Liu, Y. T., and Zhang, Y. Q. (2023). Physics-informed neural network-based petroleum reservoir simulation with sparse data using domain decomposition. Petroleum Science, 20(6), 3450-3460.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biological cybernetics, 43(1), 59-69.

Lee, D. D., and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788-791.

Lee, D., and Seung, H. S. (2000). Algorithms for non-negative matrix factorization. Advances in neural information processing systems, 13.

Lee, E., and Kim, S. (2021). Characterization of soil moisture response patterns and hillslope hydrological processes through a self-organizing map. Hydrology and Earth System Sciences, 25(11), 5733-5748.

Liu, J. J., and Liu, J. C. (2022). Integrating deep learning and logging data analytics for lithofacies classification and 3D modeling of tight sandstone reservoirs. Geoscience Frontiers, 13(1), 101311.

Lundberg, S. M., and Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

Mudunuru, M. K., Ahmmed, B., Frash, L. P., and Frijhoff, R. (2023). Deep Learning for Modeling Enhanced Geothermal Systems. In PROCEEDINGS of 48th Workshop on Geothermal Reservoir Engineering. Presented at the Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, California.

Mudunuru, M. K., Ahmmed, B., Rau, E., Vesselinov, V. V., and Karra, S. (2023). Machine learning for geothermal resource exploration in the Tularosa Basin, New Mexico. Energies, 16(7), 3098.

Mudunuru, M. K., Vesselinov, V. V., and Ahmmed, B. (2022). GeoThermalCloud: Machine Learning for Geothermal Resource Exploration. Journal of Machine Learning for Modeling and Computing, 3(4).

Nogales, A., Guadalupe, D., and Tejedor, Á. J. G. (2025). Self-organizing maps as a way to evaluate optimal strategies for balancing binary class distributions: a methodological approach.

Pham, T. S., Truong, H. M., and Pham, T. B. (2017). Application of self organizing map in construction, geology and petroleum industry. Science and Technology Development Journal, 20(K4), 30-38.

Plaksina, T., White, C., Nunn, J., and Gray, T. (2011). Effects of coupled convection and Co2 injection in stimulation of geopressured geothermal reservoirs. In 36th Workshop on Geothermal Reservoir Engineering, Stanford University (pp. 146-154).

Siler, D. L., and Pepin, J. D. (2021). 3-D geologic controls of hydrothermal fluid flow at Brady geothermal field, Nevada, USA. Geothermics, 94, 102112.

Siler, D. L., Pepin, J. D., Vesselinov, V. V., Mudunuru, M. K., and Ahmmed, B. (2021). Machine learning to identify geologic factors associated with production in geothermal fields: A case-study using 3D geologic data, Brady geothermal field, Nevada. Geothermal Energy, 9, 1-17.

Siler, D. L., Pepin, J. D., Vesselinov, V. V., Mudunuru, M. K., and Ahmmed, B. (2021). Machine learning to identify geologic factors associated with production in geothermal fields: a case-study using 3D geologic data, Brady geothermal field, Nevada. Geothermal Energy, 9, 1-17.

Tewari, S., and Dwivedi, U. D. (2019). Ensemble-based big data analytics of lithofacies for automatic development of petroleum reservoirs. Computers and Industrial Engineering, 128, 937-947.

Vesselinov, V. V., Ahmmed, B., Mudunuru, M. K., Pepin, J. D., Burns, E. R., Siler, D. L., and Middleton, R. S. (2022). Discovering hidden geothermal signatures using non-negative matrix factorization with customized k-means clustering. Geothermics, 106, 102576.

Vesselinov, V. V., Mudunuru, M. K., Ahmmed, B., Karra, S., and Middleton, R. S. (2020). Discovering signatures of hidden geothermal resources based on unsupervised learning (No. LA-UR-20-21030). Los Alamos National Laboratory (LANL), Los Alamos, NM (United States).

Vesselinov, V. V., Mudunuru, M. K., Karra, S., O'Malley, D., and Alexandrov, B. S. (2019). Unsupervised machine learning based on non-negative tensor factorization for analyzing reactive-mixing. Journal of Computational Physics, 395, 85-104.

Wang, S., Chen, Z., and Chen, S. (2019). Applicability of deep neural networks on production forecasting in Bakken shale reservoirs. Journal of Petroleum Science and Engineering, 179, 112-125.

Wang, Z., Tang, H., Cai, H., Hou, Y., Shi, H., Li, J., and Feng, Y. (2022). Production prediction and main controlling factors in a highly heterogeneous sandstone reservoir: Analysis on the basis of machine learning. Energy Science and Engineering, 10(12), 4674-4693.

Yin, H. (2008). The self-organizing maps: background, theories, extensions and applications. In Computational intelligence: A compendium (pp. 715-762). Berlin, Heidelberg: Springer Berlin Heidelberg.

Zhang, P., Jia, Y., and Shang, Y. (2022). Research and application of XGBoost in imbalanced data. International Journal of Distributed Sensor Networks, 18(6), 15501329221106935.